ASTROBIOLOGY
Volume 14, Number 6, 2014

DOI: 10.1089/ast.2013.1127

Data Management in Astrobiology: Challenges and Opportunities for an Interdisciplinary Community

Arsev Umur Aydinoglu,[1] Todd Suomela,[2] and Jim Malone[2]

Abstract

Data management and sharing are growing concerns for scientists and funding organizations throughout the world. Funding organizations are implementing requirements for data management plans, while scientists are establishing new infrastructures for data sharing. One of the difficulties is sharing data among a diverse set of research disciplines. Astrobiology is a unique community of researchers, containing over 110 different disciplines. The current study reports the results of a survey of data management practices among scientists involved in the astrobiology community and the NASA Astrobiology Institute (NAI) in particular. The survey was administered over a 2-month period in the first half of 2013. Fifteen percent of the NAI community responded ($n = 114$), and additional ($n = 80$) responses were collected from members of an astrobiology Listserv. The results of the survey show that the astrobiology community shares many of the same concerns for data sharing as other groups. The benefits of data sharing are acknowledged by many respondents, but barriers to data sharing remain, including lack of acknowledgement, citation, time, and institutional rewards. Overcoming technical, institutional, and social barriers to data sharing will be a challenge into the future. Key Words: Data management—Data sharing—Data preservation. Astrobiology 14, 451–461.

1. Introduction

Data management for researchers is a growing concern across many disciplines. The origins of this concern include bottom-up and top-down demands on resources. There are a growing number of calls by researchers themselves to share data with others in order to help test results and reproduce experiments (Carpenter, 2012; Sarewitz, 2012; Bastian, 2013). In addition, funders and policy makers are placing demands on scientists to provide data management plans with their grant applications (NSF, 2010; NASA, 2011; Bobrow, 2013). Astrobiology is experiencing these demands just like many other scientific disciplines.

The current study is the first to survey astrobiologists about their current data management practices and attitudes toward data sharing. Knowing more about current practices and attitudes will help develop future policy recommendations and allocate resources for the development of infrastructure. The survey may be used to identify current data management problems, notify funders of resource gaps, and help establish strategic partnerships with other disciplines such as information science. A further goal of the current paper is to raise the awareness of the astrobiology community regarding the challenges of data management. The introduction describes top-down and bottom-up drivers of data management, discusses some institutional mandates for data management in various disciplines, and reviews the barriers to and benefits of data sharing that have been found in the literature.

The immediate impetus to create data management plans may come from organizations such as the NASA Astrobiology Institute (NAI), which has asked research teams to include a data management plan in their proposals for the next funding cycle, which will be due in early 2014. According to the Cooperative Agreement Notice (CAN), ''the data management plan should ensure that results are fit for contemporary use and available for discovery and reuse,'' and data should be made openly available in 2 years (NAI, 2013, p 20). NAI is among a growing number of federal funding agencies in the United States that require a data management plan section in the grant proposal submissions (Bobrow, 2013), as the benefits of data reuse and sharing have been acknowledged increasingly by the scientific community and policy makers. For instance, the National Science

[1]NASA Astrobiology Institute, NASA Ames Research Center, Moffett Field, California.
[2]School of Information Sciences, University of Tennessee, Knoxville, Tennessee.

Foundation (NSF) requires a supplementary data management plan to be submitted with all proposals (NSF, 2010), the National Institutes of Health (NIH) have a similar policy since 2003 for grants $500,000 and higher (NIH, 2003), and the National Oceanic and Atmospheric Administration (NOAA, 2013) created a wiki page to support its community. NASA also has a full and open data sharing policy. According to NASA Earth Science Data & Information Policy (NASA, 2011):

> NASA promotes the full and open sharing of all data with the research and applications communities, private industry, academia, and the general public. The greater the availability of the data, the more quickly and effectively the user communities can utilize the information to address basic Earth science questions and provide the basis for developing innovative practical applications to benefit the general public.

The challenges and benefits of data management and data sharing are multiple. Numerous studies have documented the barriers to having a robust, secure, accessible, and interoperable cyberinfrastructure to keep data (PARSE Insight, 2009; Tenopir *et al.,* 2011). Barriers to sharing data are similar across many disciplines. Some of the practical barriers to data sharing include a lack of funding, accessibility of data sets, data standards, infrastructure, and time for researchers to do what is necessary to make data sharing possible (Reichman *et al.,* 2011; Enke *et al.,* 2012). Technological and human resource infrastructure is required for data to be collected, described, organized, and curated. Funding for the infrastructure to enable effective data management is likely the greatest barrier to data sharing (Berman and Cerf, 2013).

The diverse research communities in astrobiology present additional challenges. Astrobiology is quite interdisciplinary; 110 different disciplines are represented in projects conducted in 2012 (Aydinoglu *et al.,* 2013). The diversity of the community and research conducted by the community creates its own challenges since the differences in values, concepts, work flows, and practices can be quite divisive. There is also the competition aspect of the problem that inhibits ''the free exchange of ideas and data'' in the NAI community (National Research Council, 2008), which is a barrier not only to data sharing but also to collaboration.

Despite these barriers, there are many benefits of sharing data, including reanalysis of data to verify results, reinterpretation of data with alternate disciplinary approaches, maintenance of data integrity via proper preservation, avoidance of redundant data collection, and use of data as a training tool for beginning researchers (European Science Foundation, 2007; Hanson *et al.,* 2011; Reichman *et al.,* 2011; Tenopir *et al.,* 2011). The progress of science and accelerating new innovations depends on scientific findings being made available (Wallis *et al.,* 2013). This allows for confirmation or refutation of the results and is a cornerstone of the social contract within science (Vision, 2010). In addition to benefiting the scientific community, sharing of raw data has been associated with an increase in citation rates of the resultant publications. In a study of 85 cancer microarray clinical trial publications, Piwowar *et al.* (2007) found that making data publicly available resulted in a 69% increase in citations. Although causation could not be proven, this study does point to the importance of sharing data. This is a direct benefit to the authors of clinical trial publications, since citation rates can inform pay increases and promotion potential.

The benefits of data sharing are widespread and acknowledged by most scientists. In a broad survey of data-sharing trends among 1329 scientists in North America, Europe, and Asia, Tenopir *et al.* (2011) found that most were in favor of sharing data (75%) and interested in reusing data sets from others (83%) if easy to access, but less than 6% actually shared all their data with others. Participants in this survey represented several disciplines: environmental sciences and ecology, social sciences, biology, computer science/engineering, atmospheric science, and medicine. In another study published in *PLoS One* (Alsheikh-Ali *et al.,* 2011), the activity of authors of 500 papers in 50 high-impact journals was examined to determine publishing policies of underlying scientific data sets. There was wide variation in the publisher requirements, with 30% not requiring raw data submission, while the study showed that 59% of papers published in the journals with a data policy did not fully adhere to the requirements for publishing underlying data. Overall, only 47 papers (9%) included access to the full raw data online. These two interdisciplinary studies show that there is room for improvement in data management practices across scientific fields.

Recently, the U.S. Office of Science and Technology Policy (OSTP, 2013) issued a memo directing federal agencies with a budget larger than $100 million annually for R & D to make all federally funded research results (peer-reviewed publications and digital data) available to the public, industry, and the scientific community. The challenge is balancing calls for greater public access to scientific data with the constraints of current funding and institutional structures. A robust science data infrastructure is a critical issue and will require collaboration across the whole information chain of authors, research institutions, data centers, libraries, and publishers (PARSE Insight, 2009). Achieving such a system will require input from public, private, and academic stakeholders. Libraries may be one avenue toward curating and providing access to some data while moving toward a long-term sustainable economic model, and private companies may be persuaded to do the same for specialized data repositories (Berman and Cerf, 2013).

The current study is one part of a larger research project to investigate the interdisciplinary collaborative practices at NAI (the other three parts are currently in progress and cover communication behaviors, collaborative works and interdisciplinary interaction, and institutional identity). The aim of the current study is to gain a better understanding of the perceptions toward, and practices of, scientific data management (preservation, sharing, and reuse) in the astrobiology community. The study is important (i) because the nature of astrobiology requires interdisciplinary collaborative science, and data sharing is an integral part of it; (ii) because the astrobiology community consists of young researchers who can be more receptive to training in data management; and (iii) because of the recent decision by NAI to require a data management plan in proposals. The findings will benefit not only NAI management and astrobiology researchers but also the information science community working on data management, data sharing, data preservation, and cyberinfrastructure/eScience.

2. Methods

The survey instrument was based on the seminal study of Tenopir *et al.* (2011), which was designed as a baseline assessment to explore data-sharing practices of scientists. The call was open to all disciplines; and mostly researchers from environmental sciences and ecology, social sciences, life sciences, physical sciences, computer sciences, and even medicine participated, which is a good proxy for astrobiology, as astrobiology is quite diverse. The original instrument asked for demographic information, funding agency, country of employment, and professional position. These questions were retained for this survey but modified to reflect different funding agencies and the organizational structure of NAI. The authors of the present study added new questions that were based on their studies of collaboration within the astrobiology community. Questions were modified to reflect the predominant data formats and data repositories used by the astrobiology community. Demographic questions about primary NAI team and roles on multiple teams were also added. The goal was to keep questions similar to the original survey in order to facilitate potential comparisons across disciplines.

Questions about data management practices in the astrobiology community asked for types of data collected, data formats, and use of metadata standards. Questions about data backup practices asked for primary data storage locations and frequency of backups. The options for these questions were based on interviews and consultations with experienced members of the astrobiology community. For instance, an informal data management session was held at the 2012 Astrobiology Science Conference with about 30 participants. Their feedback was integrated into the survey and used in the interpretation of data.

Five sets of questions asked for attitudes, perceptions, and practices regarding data. Respondents were asked why their data should not be made available and what was their degree of satisfaction with current processes used for data management, benefits of data sharing, organizational support for data sharing, and conditions for sharing data with other researchers. Each set of questions used a 5-point Likert scale of *disagree strongly, disagree somewhat, neither agree nor disagree, agree somewhat,* and *agree strongly. No answer* was also an option.

Four postdoctoral researchers and one senior NASA researcher who works on data management were contacted to get feedback regarding the preliminary survey questionnaire. Postdoctoral researchers were selected according to their disciplinary background in order to cover the most commonly represented fields in the astrobiology community (astrobiology and engineering, physics, geosciences, and biology). The surveys were sent to the postdocs, and the feedback was received via e-mail. Afterward, a face-to-face or online meeting (30–60 min) was scheduled with postdocs to discuss the feedback in detail. Modifications were made to the survey to improve clarity and tailor the instrument to the population being sampled. (The survey is available online as supplementary material at www.liebertonline.com/ast.)

The survey was distributed via e-mail to a Listserv hosted by NAI. The Listserv has 4345 members and may be joined by anyone who wants to receive the *International Astrobiology Newsletter.* There were 750 NAI researchers included on the Listserv. The survey was available from April 22 to July 1, 2013. Two recruitment e-mails were sent on April 22 and June 28. The survey instrument was hosted on University of Tennessee servers and approved by the Institutional Review Board of the university.

After the survey was closed, 194 responses were judged complete. Of those 194 responses, 114 were from members of the NAI teams, a response rate of 15%. The other 80 responses were from people outside the NAI network. The overall response rate, including all participants from NAI and the Listserv, was 4.5%.

3. Results

3.1. Demographics

One hundred ninety-four people responded to the survey. On average, 52% of the overall respondents' reported time is spent on research, 15% on teaching, 14% on administration, and 9% on outreach activities. One hundred thirteen respondents (57%) indicated that they spent 50% or more of their time on research. In short, the sample consists of research-intensive people. Approximately three-quarters of the respondents are from academia (73.2%), followed by government (16.5%), nonprofit organizations (6%), and commercial organizations (2.5%). They are at different levels of their careers, from graduate students to senior scientists, both in academic and government scientist tracks. Primary place of employment for the participants is the United States (70%), followed by the European Union countries (14%). One-third of the respondents are female.

Of the 194 respondents, 114 are involved with the NAI teams, and 80 are not. Members from all the NAI teams, except one, responded to the survey (see Fig. 1). NAI teams are a collaboration of researchers from different levels (from senior researchers to graduate students), and their contribution is labeled accordingly. All the research-intensive roles are represented in the survey.

3.2. Data management

The importance of data sharing and preservation is widely acknowledged by the participants. Eighty-three percent of the respondents agree with the statement that ''well-maintained data helps retain data integrity''; 82% agree that ''reanalysis of data helps verify results data''; 78% agree with ''data sharing reduces redundant data.'' According to the participants, data management practices are beneficial to the scientific process itself (verify results data) and to the interdisciplinary collaborative science and the training of the next generation of researchers (see Fig. 2).

3.3. Data

The types of data the participants use are diverse (see Fig. 3). Experimental data (with some manipulation) is the most frequently used type (121), followed by observational data (no manipulation involved) and data models (99 and 74, respectively). The high number for social science data (13) and interviews (11) may be due to overrepresentation of education and outreach team members among the survey respondents. In any case, the diversity of data types used by the participants demonstrates the strength of the community

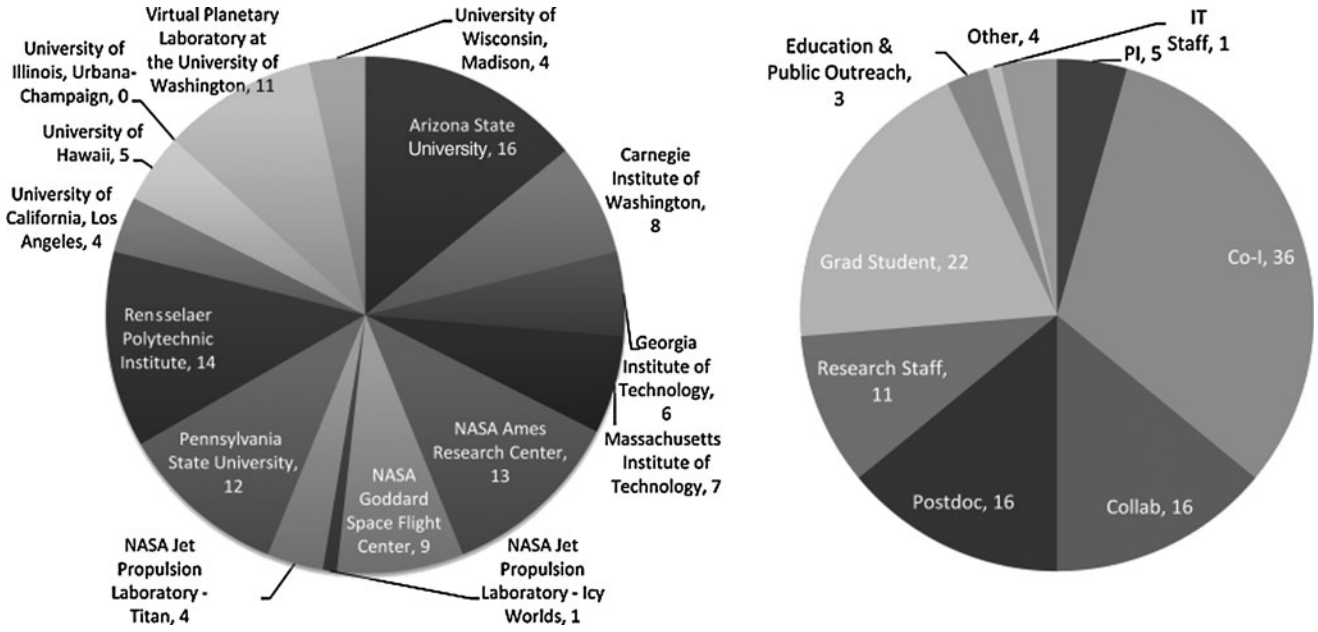

FIG. 1. Participation in the survey based on NAI teams (left); participation in the survey based on role in the NAI teams (right); $n = 114$ in both charts.

and presents a challenge if NAI wants to create a shared cyberinfrastructure or combine different repositories for astrobiology-related data.

The participants were also asked what type of data formats they use to store data. One hundred and twenty-eight participants indicated that they use spreadsheets to store data, 104 use txt files (ASCII text file—flat, rectangular, hierarchical, comma- or tab-delimited), 82 use freetext, and 70 use comma-separated values (see Fig. 4). Statistical data package formats are also commonly used (60); Matlab is the most frequent one among them, approximately three-quarters. Image as a data format is also prominent (flexible image transport system—fits—and jpg/jpeg are the most common among them).

The participants back up their data (only two people said ''never''), but the frequency of the backup varies. Nineteen percent of them said that they back up their research data immediately, 23% daily, and 20% weekly. Multiple media are used to preserve data. In addition, each participant uses more than one medium to back up data. Since frequent field visits where connectivity is limited are common among astrobiology researchers, it is not a surprise that a great majority of the respondents' first choice is portable backup (external hard drives, 151; thumb drives, 4). Internal hard

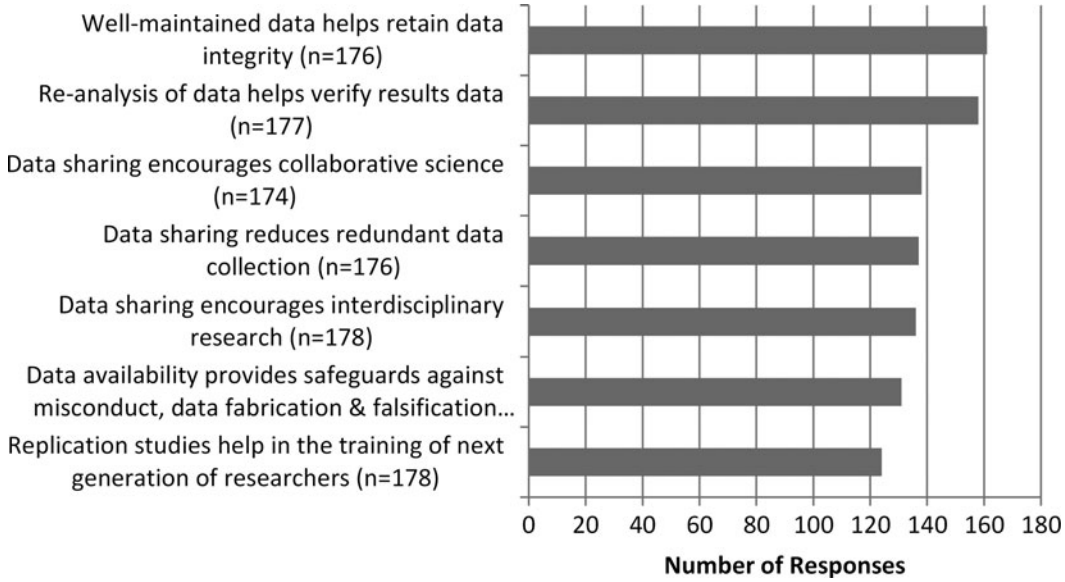

FIG. 2. Benefits of data sharing and preservation.

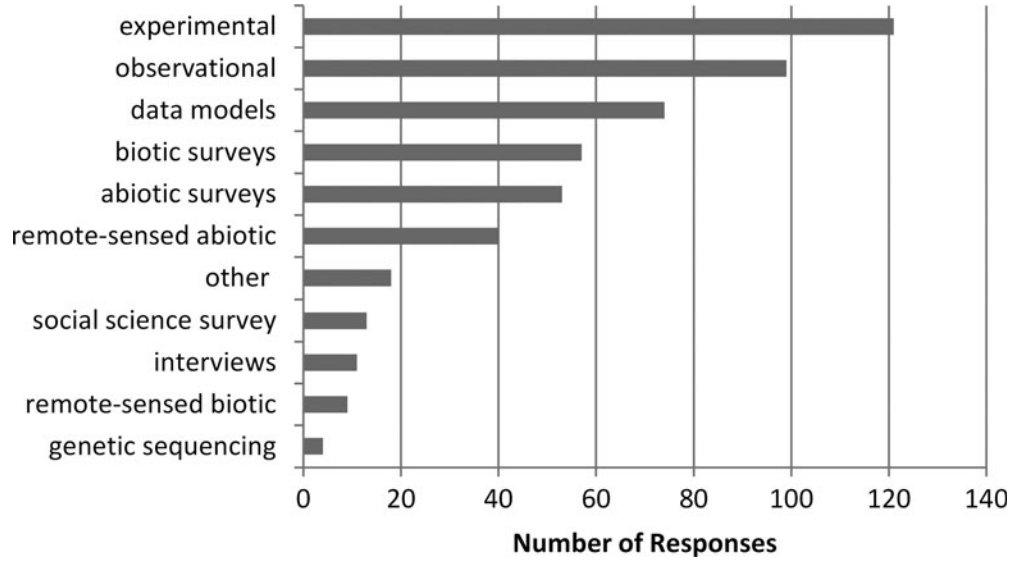

FIG. 3. Data types used ($n = 194$).

drives (68) and cloud (67) are the second choice. CD/DVDs (41) and magnetic tapes (6) are still around ($n = 194$). Only a small fraction of the participants use institutional repositories (12). The penetration of cloud-based services (or institutional repositories) is low. However, when the participants were asked where they store their research data, 148 of them responded ''on a personal computer connected to the network''; 82 ''on a computer operated by my school, company, or organization''; and 81 ''on a computer operated by my research team'' ($n = 194$). Only a limited number of them were publicly accessible, as follows: 34 ''on a public website'' and 27 ''on an open access repository.''

The diversified nature of the astrobiology community is reflected in the metadata standards preferred as well. Twenty-eight of the participants responded with ''metadata standardized within my lab,'' thirteen with ''International Standards Organization (ISO),'' and seven with ''Astronomy Visualization Metadata Standard (AVM)'' and ''Open GIS (OGIS)'' ($n = 191$). In addition, 42% of the respondents either do not capture metadata or ''don't know'' whether they capture it.

3.4. Data sharing

The participants who share their research data were asked whom they share it with. A great majority (79%) responded that they share it with other members of their research team ($n = 178$). They also share it with ''other researchers at my

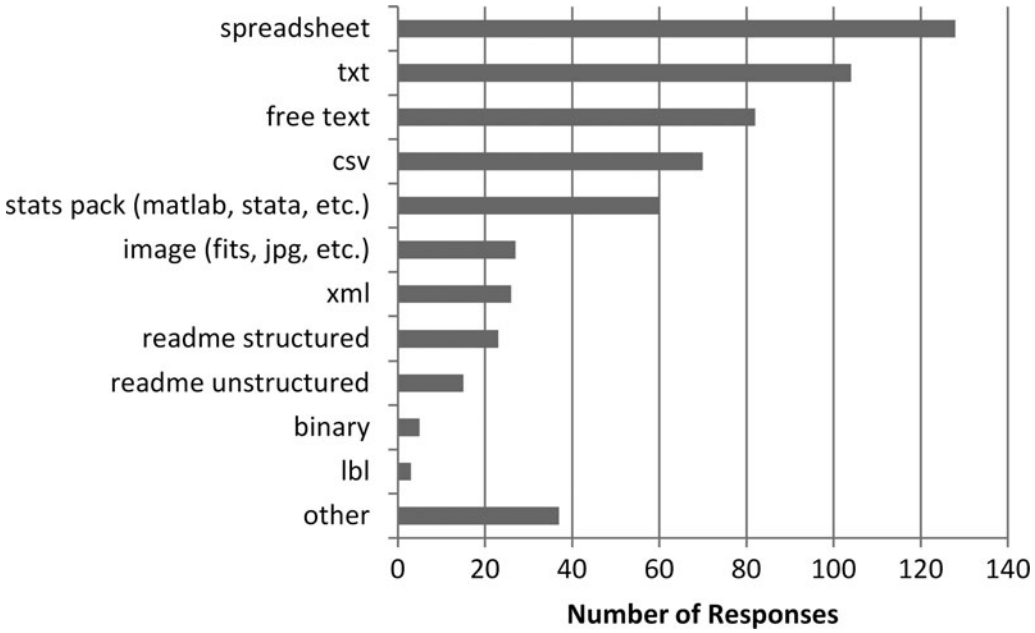

FIG. 4. Data formats ($n = 194$).

institution,'' ''other researchers in my discipline,'' and ''scientific community at large'' (90, 89, 88, respectively). Close to a quarter of the respondents (25.8%) share it with the funding agency.

The majority of the respondents store their data on a personal computer connected to the network (see Fig. 5). The next most common locations for data storage are on computers operated by local organizations such as a school or research team. Public access, on a Web site or repository, is much less common than private storage. Public Web sites, open and commercial repositories represented only a small portion of the data storage opportunities for the respondents. Figure 5 shows a bar graph of the number of participants and where they store their data. Participants could mark more than one answer.

The interdisciplinary and fragmented nature of the astrobiology community is evident in the responses to the data repositories being used. Sixty-two different repositories were named (see appendix Table A1 for the full list). NASA, ESA, and the National Center for Bioinformatics (NCBI) are in charge of most of them. The most common disciplines represented in these repositories are astrophysics, genetics, exoplanets, and earth sciences research.

Although the participants acknowledge the benefits of data sharing, there are a variety of reasons why they do not make data available to others (see Fig. 6). A little above a quarter of all the respondents think that ''people don't need them [research data]'' ($n = 194$). Another common reason is lack of time (43), funding (35), and space (32). These could all be categorized under ''lack of resources'' together with ''don't have technical skills & knowledge'' (22).

Researchers are generally satisfied with the process for collecting their research data (''agree'' 79%, $n = 180$) and with the process for searching their own data (''agree'' 73%, $n = 179$). The process for cataloging/describing research data is also satisfactory for approximately two-thirds of the participants (''agree'' 68%, $n = 179$), together with the process for storing data during the life of the project (''agree'' 70%, $n = 178$). However, less than half the participants are satisfied with the process for storing data beyond the life of the project (''agree'' 45%, $n = 179$), which indicates a data preservation issue.

Seventy-one percent of the respondents share their data with others ($n = 177$); however, only 30% of them think that ''others can access their data easily'' ($n = 180$). This difference can be explained partly by the fact that only 19% of the participants are satisfied with the tools they are using for preparing metadata ($n = 171$) and/or 43% are satisfied with the tools for preparing their documentation ($n = 176$). In addition, only 36% of the participants' organization or project has a formal established process for managing data ($n = 181$), and 39% provide the necessary tools and technical support for data management ($n = 180$). A quarter of the participants said that they are provided the necessary funds to support data management ($n = 180$), and only 16% received training on best practices for data management by their organizations and projects ($n = 180$). Lack of organizational support on research data management seems to be a problem. A participant wrote, ''Universities are trying to support mandated data sharing but do not appreciate the scale of the problem.''

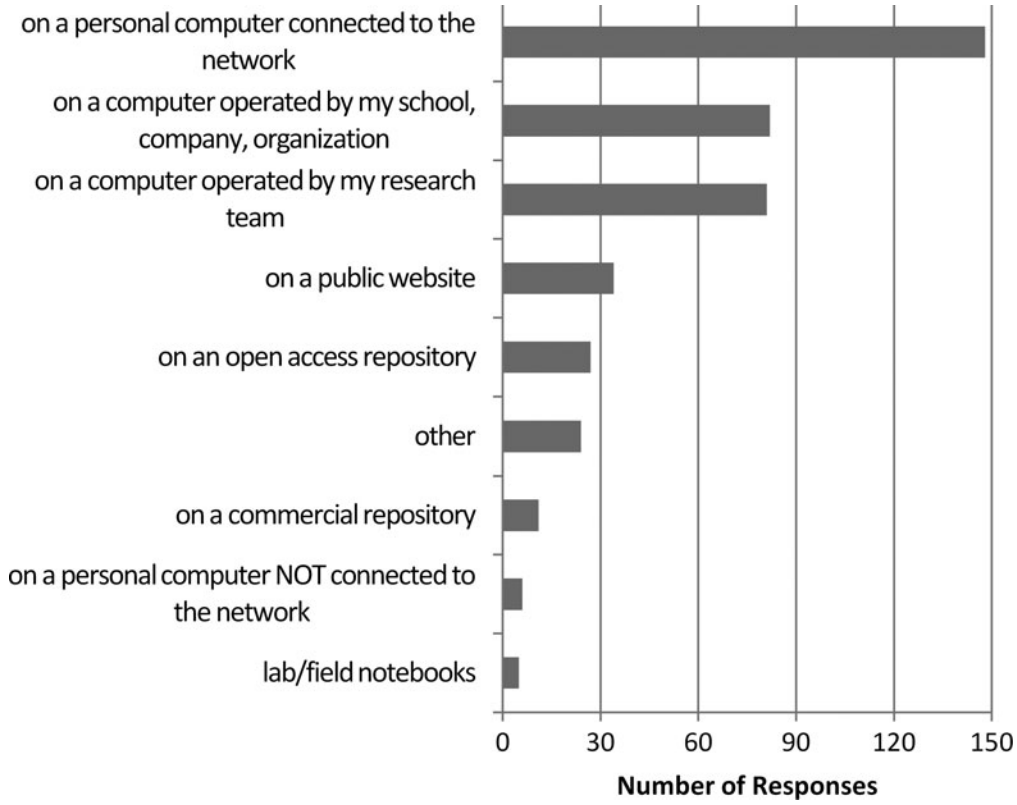

FIG. 5. Places where research data is made available to others ($n = 194$).

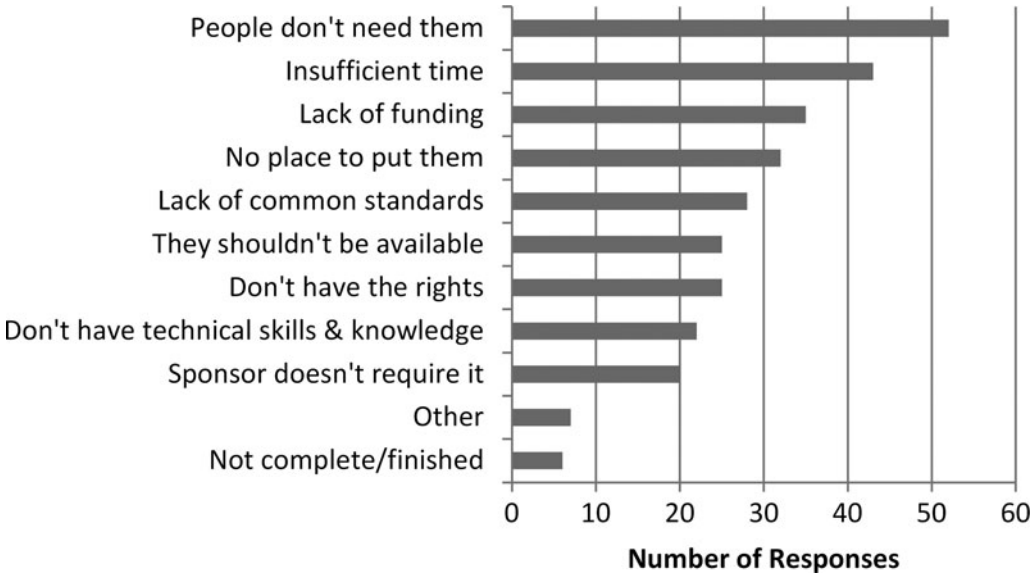

FIG. 6. Reasons for not making research data available to others.

The participants are willing to share their data if certain conditions are met (see Fig. 7), which suggests that, if an ecosystem is created that encourages these conditions, the behavior of the community can be changed. The most common conditions are as follows: formal acknowledgement of the data providers and/or funding agencies in all disseminated work making use of the data (90.6%); preliminary data should be labeled as such so that people know when data are not completely vetted (90%); formal citation of the data providers (87.60%); the opportunity to collaborate on the project (86.40%); coauthorship on publications resulting from use of the data (80.50%); and availability to the team but not outside until publication (79%). Some of these responses are highlighted by the participants in the comments section: for instance, on acknowledgement, ''making data citable is a great idea, so that those who produce them get proper credit''; on authorship, ''data sharing…should be done while respecting the right of first authorship for the team/project collecting data''; on embargo period, ''data acquired with the support of public funds…should be made available after an appropriate proprietary period.''

Only 28% of the participants said that their primary project funding agency requires them to provide a data management plan. This has been changing; NSF made this a requirement in 2010, and now NAI has for the next grant cycle. Other federal agencies are supposed to implement data management plans as well. In February 2012, the U.S. Office of Science and Technology Policy issued a memo directing all federal agencies with budgets >$100 million to share their data publicly (OSTP, 2013).

4. Discussion

One of the major drivers toward data sharing is top-down research policy. NASA Earth Science Division's policy has recognized the importance of data sharing (NASA, 2011), and NAI is asking for a data management plan for all grants submitted in the next grant cycle. Other agencies, such as NSF and NIH, have mandated data management plans as part of the funding application process (NIH, 2008; NSF, 2010). Another example is EarthCube ''data and knowledge management system'' for geosciences as a partnership between NSF's Directorate for Geosciences and the Division of Advanced Cyberinfrastructure (EarthCube, 2013). At the same time, there are bottom-up efforts by the scientific community to identify a problem and take action to deal with it. For instance, both DataONE and Data Conservancy (a cyberinfrastructure for Sloan Digital Sky Survey) consist of concerned scientists who wanted to address data management issues and started working on them years before the NSF DataNet solicitation in 2008 (NSF, 2008; Aydinoglu, 2011; Data Conservancy, 2014).

Today, the growth of top-down and bottom-up demands for data sharing represents an opportunity for scientists and funding organizations to collaborate and improve the data management infrastructure for astrobiology. The results of the current survey show that many members of the astrobiology community are willing to share their data but are frustrated by the lack of tools for creating metadata or documentation. Formal procedures are not in place, and most data sharing is ad hoc. Nor is there much infrastructure, training, or tools available for helping people perform or learn about best practices in data management.

Those in a position to make recent changes in policy should be careful about creating repositories where data is dumped but not reused because it is not vetted, easily accessible, or well defined, and supporting materials (such as analysis software, reliable error estimates associated with data, or lab notebooks) are not provided. As one participant responded, ''I am concerned that all federal research budgets will be taxed to support the release [of] reams of garbage data.'' Appropriate data quality assurance, control, discoverability, and preservation are important steps to avoid ''garbage'' data, but they are costly. To date, there have not been many mechanisms available to support such activities beyond the life of the individual project. Building data management plans that integrate such costs from the

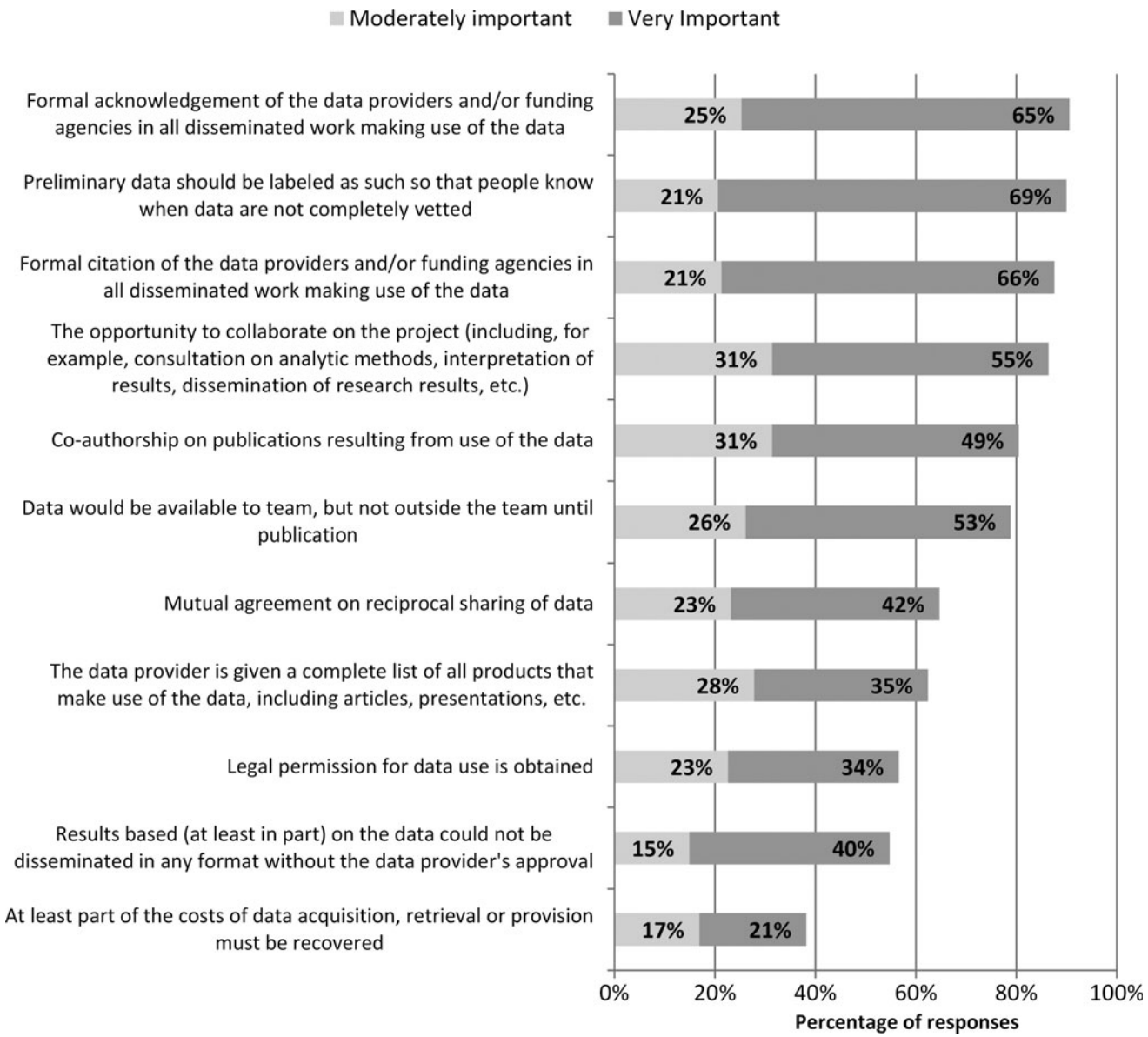

FIG. 7. Conditions for sharing data.

beginning are an important step in the right direction and a motivation for top-down changes in research policy.

The survey results show two major challenges for the astrobiology community when it comes to data management. The first is the difficulty of extending data sharing beyond isolated research teams. The second is the diversity of data and academic disciplines within the astrobiology community. The results of the current survey show that members of the astrobiology community share their data with other members of their research teams, but the sharing declines as the group of potential recipients expands beyond the local level. The survey respondents identified a number of reasons why they do not share data. Perceived lack of need was the most frequent response, followed by insufficient time, lack of funding, no place to put the data, and lack of common standards. Previous studies of other academic disciplines show lack of funding and insufficient time as the most frequent reasons for not providing electronic access to data (Tenopir *et al.,* 2011; Enke *et al.,* 2012). Providing an outlet for data sharing beyond the local level is one area where funding agencies may be able to direct the astrobiology community; thus, NAI's data management plan requirement for CAN 7 is very appropriate. In addition, there is an interest in the information science community to support research data management. Reaching out to these communities could help provide training and other synergies to the astrobiology community.

The second major challenge to sharing data within the astrobiology community is the large diversity of disciplines, organizations, and ages represented within the research community. Academia, government, nonprofit institutions, and industry are all represented within the NAI community with tens of disciplines from life sciences to earth sciences to space sciences and even social sciences. Such a diverse population presents unique challenges for data sharing and reuse such as multiple data types collected, different formats

used, numerous data repositories, and the lack of metadata standards. The wide diversity is reflected by the current survey in the number of different data types and metadata formats used by the respondents. Collaborative data is one of the keys to facilitating interdisciplinary team science, in addition to cultural norms, technologies, and mutual respect (Stokols *et al.,* 2008). Diversity of data types, formats, repositories, and metadata is a challenge for interdisciplinary science, and it is imperative that any data preservation system support standards for all relevant subject disciplines and data formats, with an ease of conversion between the different metadata standards (Tenopir *et al.,* 2011).

In addition to the challenges of data management, there are two unique characteristics of the astrobiology community that increase the likelihood that astrobiology will successfully deal with data management and sharing. They are diversity and youth. Diversity is more than just a barrier, it is also an opportunity. A grand challenge, like identifying the building blocks and environments necessary for the origin and evolution of life, requires significant levels of collaboration and interdisciplinary cooperation (Cockell, 2002; Omenn, 2006; Voytek, 2011). Scientists involved in astrobiology need to be able to use data, theories, models, and samples from many different fields because the processes are interdependent and complex (Des Marais *et al.,* 2008). Improving data management is necessary to help astrobiology answer fundamental research questions.

The relative youth of the astrobiology community presents another opportunity. Half the NAI-funded researchers received their PhD degrees in the last 12 years; furthermore, there have been over 200 graduate students in NAI teams in any given year in the last 5 years (Aydinoglu, 2013). Research has shown that ''younger people were more likely to think lack of access to data is a major impediment to progress in science and has restricted their ability to answer scientific questions'' (Tenopir *et al.,* 2011, p. 15). The data-sharing habits of the next generation of scientists are influenced by their initial training experience as graduate students and in postdoctoral assignments, and a firm understanding of this will allow faculty and information professionals the opportunity to have a positive effect on the data culture of the future (Vogeli *et al.,* 2006).

Harnessing the interdisciplinary and collaborative nature of the astrobiology research community to inform data management practices should be a goal within the community. Improving efforts to share data in astrobiology could enable a progressive and positive ripple effect into the wider science communities from which astrobiology draws its members. Funding agencies and professional organizations may provide resources for server space, training, best practices, toolkits, and more. Collaborations with information professionals may provide learning opportunities such as classes and materials for science informatics and data curation. Projects such as DataONE have demonstrated the value of interdisciplinary collaborations across computer sciences, library and information sciences, and earth sciences (DataONE, 2012). Members of Data Conservancy ''developed a model and framework for connecting data with publications, which was ultimately adapted and implemented for our [Data Conservancy] current service that supports data deposit for arXiv.org pre-prints'' (Data Conservancy, 2014) and made significant advances in cyberinfrastructure for earth, life, and social sciences domains.

The NASA Astrobiology Institute has achieved considerable success ''in initiating interdisciplinary research'' and ''developing collaborations'' (National Research Council, 2008). The experiences, accomplishments, failures, and lessons learned that have accumulated over the years put NAI in a unique position both in terms of the value it can add and the responsibility it has. The next step is to cocreate a new digital space in the era of data-intensive science to further collaboration in astrobiology and beyond through data sharing.

Acknowledgments

This study is supported by the NASA Astrobiology Institute. In addition, we would like to thank the Usability & Assessment Working Group of DataONE for making their survey instrument available to public.

Author Disclosure Statement

No competing financial interests exist.

Abbreviations

NAI, NASA Astrobiology Institute; NIH, National Institutes of Health; NSF, National Science Foundation.

References

Alsheikh-Ali, A.A., Qureshi, W., Al-Mallah, M.H., and Ioannidis, J.P. (2011) Public availability of published research data in high-impact journals. *PLoS One* 6, doi:10.1371/journal.pone.0024357.

Aydinoglu, A.U. (2011) Complex adaptive systems theory applied to virtual scientific collaborations: the case of DataONE. PhD dissertation, University of Tennessee, Knoxville, TN. Available online at http://trace.tennessee.edu/utk_graddiss/1054.

Aydinoglu, A.U. (2013) Knowledge transfer in the NASA Astrobiology Institute. In *5th Annual Research Conference,* National Organization of Research Development Professionals, Chicago, IL.

Aydinoglu, A.U., Allard, S., and Mitchell, C. (2013) Evolution of virtual research collaborations. In *27th Annual Conference of the American Evaluation Association,* American Evaluation Association, Washington DC.

Bastian, H. (2013, September 10). Opening a can of data-sharing worms. In *Absolutely Maybe: Evidence and Uncertainties about Medicine and Life, Scientific American* blog network, Nature Publishing Group, New York. Available online at http://blogs.scientificamerican.com/absolutely-maybe/2013/09/10/opening-a-can-of-data-sharing-worms.

Berman, F. and Cerf, V. (2013) Who will pay for public access to research data? *Science* 341:616–617.

Bobrow, M. (2013) Balancing privacy with public benefit. *Nature* 500:123.

Carpenter, S. (2012) Psychology's bold initiative. *Science* 335:1558–1561.

Cockell, C. (2002) Astrobiology—a new opportunity for interdisciplinary thinking. *Space Policy* 18:263–266.

Data Conservancy. (2014) *History of Data Conservancy,* Data Conservancy, Johns Hopkins University, Baltimore, MD. Available online at http://dataconservancy.org/about/history-of-dc.

DataONE. (2012) *What is DataONE?* Data Observation Network for Earth (DataONE), University of New Mexico, Albuquerque, NM. Available online at http://www.dataone.org/what-dataone.

Des Marais, D., Nuth, J.A., III, Allamandola, L.J., Boss, A.P., Farmer, J.D., Hoehler, T.M., Jakosky, B.M., Meadows, V.S., Pohorille, A., Runnegar, B., and Spormann, A.M. (2008) The NASA Astrobiology Roadmap. *Astrobiology* 8:715–730.

EarthCube. (2013) *History and Progress*. Available online at http://www.earthcube.org/page/history-and-progress.

Enke, N., Thessen, A., Bach, K., Bendix, J., Seeger, B., and Gemeinholzer, B. (2012) The user's view on biodiversity data sharing—investigating facts of acceptance and requirements to realize a sustainable use of research data. *Ecol Inform* 11:25–33.

European Science Foundation. (2007) *Shared Responsibilities in Sharing Research Data: Policies and Partnerships,* European Science Foundation, Strasbourg, France. Available online at http://www.esf.org/fileadmin/Public_documents/Publications/SharingData_01.pdf.

Hanson, B., Sugden, A., and Alberts, B. (2011) Making data maximally available. *Science* 331:649.

NAI. (2013) *Cooperative Agreement Notice: NASA Astrobiology Institute Cycle 7,* NASA, Washington, DC. Available online at https://nspires.nasaprs.com/external/solicitations/summary.do?method=init&solId={DF07566F-A000-4BC7-1116-34747BD49A7F}&path=closedPast.

NASA. (2011) *Data & Information Policy,* NASA, Washington, DC. Available online at http://science.nasa.gov/earth-science/earth-science-data/data-information-policy.

National Research Council. (2008) *Assessment of the NASA Astrobiology Institute,* The National Academies Press, Washington, DC.

NIH. (2003) *NIH Data Sharing Policy and Implementation Guide,* National Institutes of Health (NIH), Bethesda, MD. Available online at http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm.

NIH. (2008) *Revised Policy on Enhancing Public Access to Archived Publications Resulting from NIH-Funded Research,* National Institutes of Health (NIH), Bethesda, MD. Available online at http://grants.nih.gov/grants/guide/notice-files/NOT-OD-08-033.html.

NOAA. (2013) *Category: Data Management Plans,* NOAA Environmental Data Management Wiki, National Oceanic and Atmospheric Administration (NOAA), Washington, DC. Available online at https://geo-ide.noaa.gov/wiki/index.php?title=Category:Data_Management_Plans.

NSF. (2008) *Sustainable Digital Data Preservation and Access Network Partners (DataNet),* National Science Foundation (NSF), Arlington, VA. Available online at http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm.

NSF. (2010) *Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans,* Press Release 10-777, National Science Foundation (NSF), Arlington, VA. Available online at http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928.

Omenn, G.S. (2006) Grand challenges and great opportunities in science, technology, and public policy. *Science* 314:1696–1704.

OSTP. (2013) *Increasing Access to the Results of Federally Funded Scientific Research*. Public access memorandum from the Office of Science and Technology Policy, Washington, DC. Available online at www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

PARSE Insight. (2009) *Insight into Issues of Permanent Access to the Records of Science in Europe.* Available online at http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf.

Piwowar, H.A., Day, R.S., and Fridsma, D.B. (2007) Sharing detailed research data is associated with increased citation rate. *PLoS One* 2, doi:10.1371/journal.pone.0000308

Reichman, O.J., Jones, M.B., and Schildhauer, M.P. (2011) Challenges and opportunities of open data in ecology. *Science* 331:703–705.

Sarewitz, D. (2012) Beware the creeping cracks of bias. *Nature* 485:149.

Stokols, D., Misra, S., Mooser, R.P., Hall, K.L., and Taylor, B.K. (2008) The ecology of team science: understanding contextual influences on transdisciplinary collaboration. *Am J Prev Med* 35:S96–S115.

Tenopir, C., Allard, S., Douglass, K.L., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M., and Frame, M. (2011) Data sharing by scientists: practices and perceptions. *PLoS One* 6, doi:10.1371/journal.pone.0021101.

Vision, T.J. (2010) Open data and the social contract of scientific publishing. *BioScience* 60:330–331.

Vogeli, C., Yucel, R., Bendavid, E., Jones, L.M., Anderson, M.S., Louis, K.S., and Campbell, E.G. (2006) Data withholding and the next generation of scientists: results of a national survey. *Acad Med* 81:128–136.

Voytek, M.A. (2011) Greatest hits and grand challenges in astrobiology [abstract 2687]. In *American Association for the Advancement of Science (AAAS) Annual Meeting,* American Association for the Advancement of Science, Washington, DC. Available online at http://aaas.confex.com/aaas/2011/webprogram/Paper2687.html.

Wallis, J.C., Rolando, E., and Borgman, C.L. (2013) If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS One* 8, doi:10.1371/journal.pone.0067332.

Address correspondence to:
Arsev Umur Aydinoglu
NASA Astrobiology Institute
NASA Ames Research Center
MS:247-6
Moffett Field, CA 94043

E-mail: arsevu@gmail.com

Submitted 6 December 2013
Accepted 10 March 2014

(*Appendix follows* →)

Table A1. List of Data Repositories

ADS	GERM	NED exoplanet archives
Arxiv	GitHub	NExSci
Astrophysical Journal	Greengenes	NRAO Archives
ATNF Archives	Helix training	Orcid
CFOP	Herschel Observatory	PahDB
CONCIENCIA	HITRAN	PANGAEA
CSIC	Hubble Space Telescope Data Archive	PubChem
DOE-JGI	IPAC	R-DBMS
Earth Reference	IRSA at IPAC	SedDB
EPA	LAMDA	Seed
ESA PSA	Local Servers	Simbad
ESA USOC	Mars Orbital Data Explorer	Skyview
ESAC	MAST Archive at STScl	SRA
ESO Data Archive	MAST Scholarsphere Journal Archives	Subaru Archive
ESRIN	MG-RAST	The Cancer Genome Atlas
Exoplanet Encyclopedia	MSL	USGS
FAR-DEEP	Nanohub	VAMPS
FLWO	NASA	Vizier Hubble Space Telescope archive
GATOR for IPAC	NASA Exoplanet Archive	VPL Kepler
Genbank	NASA Planetary Data System	XSED
GEO	NCBI	